DOCUMENT RESUME

ED 228 323 TM 830 257

AUTHOR Mislevy, Robert J.; Rieser, Mark R.
TITLE Item Response Methods for Educational Assessment.
PUB DATE Apr 83
NOTE 19p.; Paper presented at the Annual Meeting of the American Educational Research Association (67th, Montreal, Quebec, April 11-15, 1983).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Comparative Analysis; Data Collection; *Educational Assessment; Item Banks; *Item Sampling; *Latent Trait Theory; Mathematical Models; *Measurement Techniques; Psychometrics; *Research Needs
IDENTIFIERS *Multiple Matrix Sampling

ABSTRACT

Multiple matrix sampling (MMS) theory indicates how data may be gathered to most efficiently convey information about levels of attainment in a population, but standard analyses of these data require random sampling of items from a fixed pool of items. This assumption proscribes the retirement of flawed or obsolete items from the pool as well as the addition of new items if comparability of estimates over time is to be maintained. Item response theory (IRT) permits measurement on an invariant scale without requiring that items be sampled at random from a fixed population, but standard IRT procedures cannot be used with the sparse data (at the level of subjects) that characterizes the most efficient multiple matrix sampling designs. This paper considers the merger of MMS theory and IRT theory. Three approaches (subject-level model, subject-level estimates; subject-level model, group-level estimates; and group-level model, group-level estimates) by which IRT can be applied in the assessment setting are described and compared. (Author/PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED228323

ITEM RESPONSE METHODS FOR EDUCATIONAL ASSESSMENT

Robert J. Mislevy

National Opinion Research Center

Mark R. Rieser

Indiana University

Abstract

Multiple matrix sampling (MMS) theory indicates how data may be gathered to most efficiently convey information about levels of attainment in a population, but standard analyses of these data require random sampling of items from a fixed pool of items. This assumption proscribes the retirement of flawed or obsolete items from the pool as well as the addition of new items if comparability of estimates over time is to be maintained. Item response theory permits measurement on an invariant scale without requiring that items be sampled at random from a fixed population, but standard IRT procedures cannot be used with the sparse data (at the level of subjects) that characterizes the most efficient multiple matrix sampling designs. This paper considers the merger of MMS theory and IRT theory. Three approaches by which IRT can be applied in the assessment setting are described and compared.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

R. J. Mislevy

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

TM 830 857

Presented at the annual meeting of the AERA, Montreal, April 1983

Introduction

Educational assessment differs from its close cousin educational measurement by its focus on the levels of attainment of groups rather than of individuals. In light of its distinct purposes, it is not surprising that methods and strategies that have proved effective in the setting of individual measurement are less than optimal in the setting of assessment. The purpose of this paper is to explore procedures under which researchers may enjoy simultaneously the advantages the assessment innovation multiple-matrix sampling and the measurement innovation item response theory.

Multiple Matrix Sampling Designs

Consider the following assessment problem: for the purposes of monitoring an educational program, it is desired to know the percent of correct responses that would be observed if all ten thousand of the pupils in a given school district were to respond to all of the one hundred test items that comprise a certain skill domain. One way to obtain this information is to administer a hundred item test to every pupil and simply take the average percent-correct score. A more economical approach is available if it is just the population average that is important; testing only a random sample of the pupils can provide sufficiently precise

estimates of the population average at a fraction of the cost. This sampling design is shown as Figure 1a.

Even more economical is a design that samples items as well as pupils. Testing time is reduced if each sampled subject is administered only a sample of the items. Figure 1b illustrates a matrix sample of this type. A multiple matrix sampling design, shown as Figure 1c, presents different samples of items to different samples of pupils. Sirotnik and Wellington's (1977) formulas for estimating means and accompanying standard errors from multiple matrix samples show that even though less is known about each individual pupil in the sample, such designs can provide more precise estimates of population attributes for a fixed number of observed responses. In fact, the most efficient design for estimating the population mean administers each item to a different sample of pupils (Figure 1d); each pupil in the sample, then, responds to only one test item (Lord, 1962). While designs that administer enough items to each sampled pupil to permit measurement of his attainment yield population estimates of decreasing efficiency, the maximally efficient design for assessment is virtually useless for measuring the attainment of specific individuals.

Practical considerations usually prevent the use of this maximally-efficient design in practice; the costs of sampling, locating, and testing generally make it more economical to administer more than one item to each sampled subject. A good compromise is to administer each sampled subject items from a

number of different skill areas, but only a few items from each. In this manner it is possible to measure population attainment in a broad spectrum of skill areas without precisely measuring individuals in any of them, thereby enjoying much of the efficiency possible from multiple matrix sampling designs.

Perhaps the greatest drawback to the direct applications of multiple matrix sampling described above is that the resulting estimate of the population average has meaning only with respect to the specific population of items from which samples were randomly drawn. It would not be possible to retire flawed or obsolete items from the collection nor to add new items without destroying the comparability of population average estimates at different points in time. As these options are highly desirable for applied work, a more flexible model is required, one in which the relationships of specific items to attainment are explicitly taken into account and attainment can be measured on a single well-defined scale even as the item domain evolves over time.

Item Response Theory

So-called "item invariant" scoring is the hallmark of item response theory (IRT). In its broadest sense, item response theory posits models in which the probability that a person will make a particular response to a given item is modelled as a function of parameters that characterize the person and parameters that characterize the item. Once items have been calibrated (i.e., their parameters have been estimated), it is possible to estimate

an individual's parameters on the basis of his responses to any subset of the items, selected randomly or not. Moreover, the person parameters can be combined with item parameters of any item--those that the person hasn't been administered as well as those that he has--to produce estimates of the probabilities of given responses. This interpretation of person parameters in IRT models is an example of "criterion-referencing" of scores.

For the most part, implementation of IRT to date has taken place in the setting of individual measurement. Indeed, estimates of individuals' attainment or aptitude levels are the desired end product of nearly all current applications of IRT. Popular methods of calibrating items and measuring individuals typically assume that enough responses have been solicited from each sampled subject to permit a reasonable estimate of his ability. It is crucial to note, however, that this is not an assumption of IRT in general but only of a specific collection of algorithms designed for a particular purpose, namely, individual measurement. These methods constitute only one of three possible approaches by which IRT can be introduced in the assessment setting.

There are two dimensions along which an application of IRT to the assessment setting might vary: the level at which a model is defined and the level at which attainment estimates are produced. The plausible combinations are these:

1. Subject-level model, subject-level estimates
2. Subject-level model, group-level estimates
3. Group-level model, group-level estimates.

The following sections consider in turn an approach based on each.

Subject-Level Model, Subject-Level Estimates

The first approach we consider borrows most directly from the technology of individual measurement. First, an IRT model is defined at the level of subjects. This means that the basic IRT model contains, in addition to one or more parameters for each item, one or more parameters that characterize the level of attainment of an individual subject. These parameters are combined in an expression for the probability of a given response to a given item from a given individual. Second, estimates of individuals' parameters are to be produced. This implies that each sampled subject must be administered enough items from any skill domain in which he is to be measured to permit a stable estimate of his ability--usually at least fifteen to twenty items. To estimate a population parameter of interest, say the mean, one estimates the attainment of each sampled subject then takes the average. This approach has been implemented in large scale assessments in the United Kingdom and in Los Angeles county in the United States.

There are certain advantages to this approach. Predominant among them is the generic IRT advantage that estimates of attainment take the form of content-referenced scores on a scale which, in theory, remains invariant even as the item pool evolves over time. Other advantages include the availablity of statistical procedures and computer programs that have been developed and

refined in the setting of individual measurement, and the opportunity for further analyses of subject-level estimates after their primary purpose of providing population-level information has been achieved.

There are also disadvantages. First, the administration of the several items necessary for subject-level IRT estimates proscribes the use of the more efficient multiple-matrix sampling designs. The decline in efficiency as more responses are solicited from each subject is dramatic. Pandey and Carlson (1976), for example, document a 700% increase in the estimation error of the mean from two designs that gather the same number of responses, from a design with a twenty items per subject to a design with a hundred items per subject. Second, because it seems wasteful to administer each subject a relatively large number of highly-related items, there is a tendency to measure skill in only a few broadly-defined domains. This practice not only restricts tha range and detail of information that can be provided for the analysis of curricular effects, but threatens the assumption of unidimensionality upon which most IRT models are based (Goldstein, 1980). The consequences of such violation may take the form of "item parameter drift," wherein estimates of an item's parameters from data at different points in time differ substantially. The content-referencing of scores and, to a lesser extent, the comparability of scores over time, is degraded to the degree that item parameter drift is present.

Subject-Level Model, Group-Level Estimates

A second approach maintains the subject-level definition of an IRT model, but bypasses the estimation of scores for sampled individuals. Instead, population parameters are estimated directly from observed response patterns by methods such as marginal maximum likelihood (Mislevy, 1983a). The basic idea is this: Let $\underline{x}_i$ be the response pattern of sampled subject i to a set of calibrated items. The item response model provides an expression for the probability of this pattern if the attainment level θ were known, say $p(\underline{x}_i|\theta)$. Let $g(\theta|\underline{\phi})$ be the density function of the distribution of attainment in the population, a function of known form but depending on the unknown values of the population parameters $\underline{\phi}$. The marginal maximum likelihood estimates of $\underline{\phi}$ are the values that maximize the probability of the observations, i.e.,

$$L = \prod_i \int_\theta p(\underline{x}_i|\theta)\, g(\theta|\underline{\phi})\, d\theta$$

An approach of this type has been employed by the International Association for the Evaluation of Educational Achievement in analyses of the Second International Study of Mathematics.

The advantage of this approach over an approach that computes subject-level estimates is the freedom to use multiple matrix sampling designs at any desired level of sampling intensity. At one extreme, this approach can applied to data from even the maximally-efficient MMS design once items have been calibrated (although at least four items per subject are desirable for item

calibration, via another marginal maximum likelihood procedure such as that described by Bock and Aitkin [1981]); at the other extreme, the same approach can be applied when indefinitely many items per scale are administered to each subject. By multivariate extensions of the basic equation shown above, it is also possible to estimate directly the relationships among skill areas or between skill areas and other demographic or educational solicited along with performance data.

The disadvantage of this approach is the requirement of statistical procedures that are more sophisticated and less familiar to the applied educational researcher. Even the estimation and subsequent analysis of IRT scores for individuals--techniques already inaccessible to many practitioners--prove inadequate for the more highly efficient designs. A tradeoff thus exists between (a) facilitating secondary research at the expense of efficiency by administering several items per subject per scale, and (b) obtaining efficient estimates of population parameters through intense item sampling, at the expense of reliance upon sophisticated analytical procedures.

Group-Level Model, Group-Level Estimates

The final approach we consider departs radically from IRT as it is practiced in the setting of individual measurement, by defining an IRT model at the level of groups of subjects (i.e., subpopulations) rather than at the level of individual subjects. Such a model contains only parameters for items and for the mean level

of attainment in groups. The basic IRT equation here does not model the probability of a given response to a given item from a particular subject, but of a subject selected at random from the group (Note 1).

This approach was introduced by Bock (1976). Since then models of this type have been implemented by the California Assessment Program (CAP), with schools as the salient groups, and by Reiser (1980) with data from the National Assessment of Educational Progress (NAEP), with the cells of the NAEP demographic design as the salient groups. These two applications represent variations on a theme: in the CAP model, a unique score is estimated for each school; in the NAEP model, the scores for the cells of the design are expressed as functions of some smaller number of main effects and selected interactions among the factors of the design.

An essential requirement of this approach is the use of the maximally-efficient MMS design, in which each sampled subject responds to only one item from any given scale. This design guarantees satisfaction of the IRT assumption of local independence in the group level model; that is, all the responses from a group may be considered independent conditional on the mean level of attainment in the group. This feature of the model entails both its advantages and its disadvantages.

The advantages of the approach are an almost spectacular efficiency and economy. Efficiency in the estimation of group means is guaranteed by MMS theory; no other design can produce estimates of the mean with more precision from the same number

of observations (see Note 2). Economy results from the ability to perform all item calibration and scale-score estimation from group-level summaries of the number of attempts and correct responses observed in the group to each item. CAP, for example, is able to perform all Grade 3 IRT analysis on a file containing data from about 4500 schools rather than a file containing responses patterns from over 200,000 pupils.

The disadvantages of a group-level IRT model stem from limitations inherent in the specification of salient groups. While scale score results from the CAP model can be aggregated upward from schools to districts, counties, and the state as a whole, no rigorous procedures are available to break results down into levels lower than schools, such as by classrooms. Also, groups must be large enough to support estimation procedures. In Reiser's "fixed-effects" model for NAEP, the number of demographic factors and their numbers of levels determines the number of cells in the design, which are the salient groups in the model; too many factors would mean too many cells, eventually with expectations of fewer than one observation per cell. And finally, once such a model has been defined and especially after summary files have been produced, it is difficult to expand the model by adding new factors or new levels of detail to existing factors.

Conclusion

Multiple matrix sampling (MMS) theory and item response theory (IRT) have been two of the most important innovations in

educational statistics over the past two decades. Until recently, however, researchers interested in assessment have not been able to enjoy the benefits of both advances simultaneously. Now, the wedding of the flexibilities of IRT with the efficiencies of MMS offers great promise to meet the challenges facing assessment in this country and others.

Practical application of IRT to the assessment setting requires careful thought, however, to avoid the problems that would occur if IRT methods developed expressly for measuring individuals were applied indiscriminantly; new approaches are often preferable in the new setting. By describing and comparing three approaches, this paper will help users avoid some of the pitfalls.

Notes

1. Under restrictive assumptions, simple relationships exist between an underlying subject-level IRT model and its associated item parameters and a group-level IRT model and its item parameters (Mislevy, 1983b).

2. In practice, it is desirable to administer to each sampled subject one item from each of a number of scales. Such a design allows efficient estimation of attainment across a range of domains with relatively narrow definitions, the errors of estimation will in general have nonzero correlations, which may be estimated by a method described in Mislevy and Bock (1981).

Appendix

Assessing the Payoff of Multiple-Matrix Sampling Designs

Formulas for estimating moments of the distribution of number correct and percent correct scores from multiple matrix sampling (MMS) data, along with formulas for estimation error variation, are found in Sirotnik and Wellington (1977). With a few simplifying assumptions, however, it is possible to approximate a key value, namely the standard error of the mean, with a much simpler formula. The basic idea is to view pupils as clusters and multiple responses from pupils as clustered samples of observations. We shall assume an infinite population of pupils and an infinite population of equally-difficult items. Samples of m items each constitute "test forms;" each sampled subject is administered one randomly selected test form; all test forms are administered to the same number of randomly selected pupils. Under these assumptions, the squared standard error of the mean is obtained as

$$E(\bar{x}-\mu) = \frac{\sigma^2}{nm}\,[1+(m-1)\rho],$$

where

n is the total number of sampled pupils,

m is the number of items per pupil (or, items per test form),

μ is the expected item response over all pupils and items,

$\bar{x}$ is the observed mean item response, or $(nm)^{-1}\sum_i \sum_j x_{ij}$, with

$x_{i,j}$ being the response of pupil i to item j,

σ^2 is the total variance, given by $E(x_{i,j}-\mu)^2$, and

ρ is the pupil intercorrelation, given by $E(x_{i,j}-\mu)(x_{i,k}-\mu)/\sigma^2$ for any items j and k.

ρ is assumed to be .5 in the examples that follow. This value is typical of standardized tests of educational achievement, and higher values might be expected in the more narrowly-defined domains that are suited to assessment. The examples will tend, therefore, to underestimate the gains in efficiency that can be obtained by more intense item sampling.

Table 1 shows the relative efficiency of several MMS designs that solicit the same number of total responses, in terms of the squared standard error of the mean. Decreasing efficiency is observed as the number of items per pupil increases. Compared with the maximally efficient design that solicits one response per pupil, only 12% as much information about the population mean is obtained in a design that solicits sixteen items per pupil, a rather minimal number of responses if the computation of scores for individuals is contemplated.

The maximally efficient design is not necessarily the best selection in a given application, however, since the costs of sampling, locating pupils, and testing pupils will increase minimally as the number of items per pupil increases. As the statistical efficiency of MMS designs increases, therefore, the costs of administration generally increase as well, and a balance must be struck to provide optimal efficiency and economy. A

chart like Table 2 proves useful in constructing a design. Table 2 shows the rate of increase in information about the population mean as the number of responses per pupil increases, for a fixed number of sampled pupils. It is clear that the rate of increase in information does not keep pace with the increase in the number of responses per pupil. In fact, under the assumptions upon which the table is based, the relative efficiency of any design in the sequence over the maximally efficient design cannot exceed $1/\rho$.

References

Bock, R.D. Basic issues in the measurement of change. In D.N.M. DeGruijter and L.J. van der Kamp (Eds.), Advances in Psychological and Educational Measurement. London: Wiley, 1976.

Bock, R.D., and Aitkin, M. Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm. Psychometrika, 1981, 46, 443-459.

Goldstein, H. Dimensionality, bias, independence, and measurement scale problems in latent trait test score models. British Journal of Mathematical and Statistical Psychology, 1980, 33, 234-246.

Lord, F.M. Estimating norms by item sampling. Educational and Psychological Measurement, 1962, 22, 259-267.

Mislevy, R.J. Characterizations of latent distributions. Manuscript submitted for publication, 1983a.

Mislevy, R.J. Item response models for grouped data. Manuscript submitted for publication, 1983b.

Mislevy, R.J. and Bock, R.D. Estimation of CAP scale scores. Technical Report #101. Chicago: International Educational Services, 1981.

Pandey, T.N., and Carlson, D. Assessing the payoffs in the estimation of the mean using multiple matrix sampling designs. In D.N.M. DeGruijter and L.J. van der Kamp (Eds.), Advances in Psychological and Educational Measurement. London: Wiley, 1976.

Reiser, M.R. A Latent Trait Model for Group Effects. Doctoral dissertation, University of Chicago, 1980.

Sirotnik, K., and Wellington, R. Incidence sampling: An integrated theory for "matrix sampling." Journal of Educational Measurement, 1977, 14, 343-399.

TABLE 1

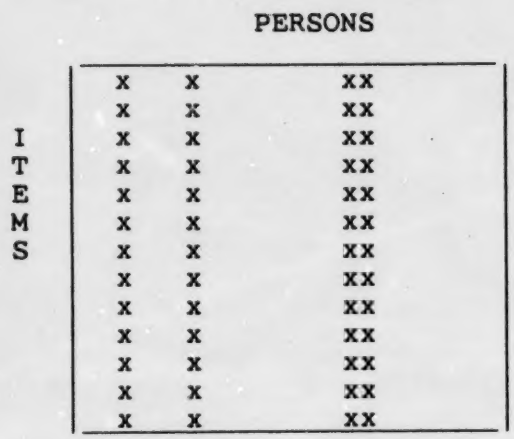
RELATIVE EFFICIENCY OF SOME SAMPLING DESIGNS WITH THE SAME TOTAL NUMBER OF RESPONSES

NUMBER OF PUPILS PER TEST FORM	NUMBER OF TEST FORMS	NUMBER OF ITEMS PER TEST FORM	TOTAL NUMBER OF RESPONSES	RELATIVE INFORMATION
100	16	1	1600	100%
100	8	2	1600	67%
100	4	4	1600	40%
100	2	8	1600	22%
100	1	16	1600	12%

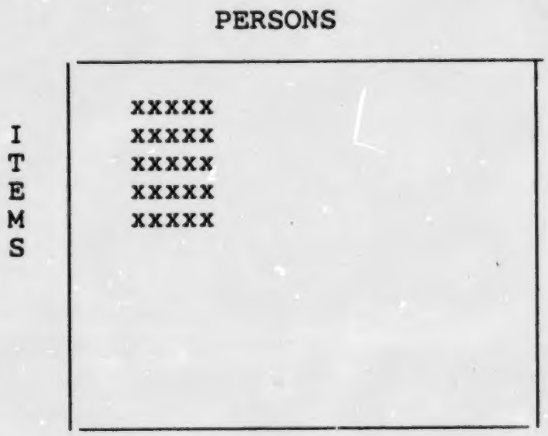
TABLE 2

RELATIVE EFFICIENCY OF SOME SAMPLING DESIGNS WITH THE SAME NUMBER OF SAMPLED PUPILS

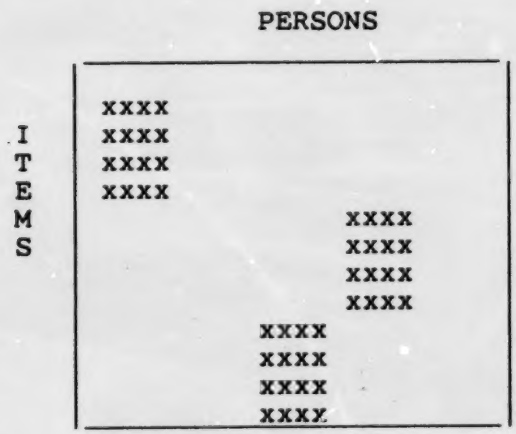
NUMBER OF PUPILS PER TEST FORM	NUMBER OF TEST FORMS	NUMBER OF ITEMS PER TEST FORM	TOTAL NUMBER OF RESPONSES	RELATIVE INFORMATION
100	16	1	1600	100%
100	16	2	3200	133%
100	16	4	6400	160%
100	16	8	12800	178%
100	16	16	25600	188%



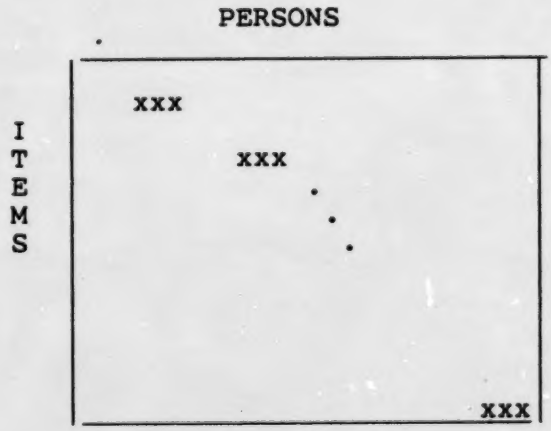
a) Subject Sampling



b) A Matrix Sample



c) A Multiple Matrix Sample



b) A Maximally-Efficient Multiple Matrix Sample